Erschienen in: Schwitalla, Johannes/Wegstein, Werner (Hrsg.): Korpuslinguistik deutsch: synchron – diachron – kontrastiv. Würzburger Kolloquium 2003. - Tübingen: Niemeyer, 2005. S. 71-80.

Werner Kallmeyer / Wilfried Schütte

Der Umgang mit Gesprächskorpora am IDS Mannheim: Die Recherche in der COSMAS-II-Gesprächsdatenbank

Gesprächsaufnahmen zu transkribieren ist eine aufwändige Arbeit. Dementsprechend langsam geht der Aufbau größerer Gesprächskorpora voran im Verhältnis zu dem schriftlicher Korpora. Das für nationale Referenzkorpora empfohlene Minimum von 10% Daten gesprochener Sprache ist angesichts des Tempos, mit dem die schriftlichen Korpora wachsen, kaum zu erreichen. Das gilt insbesondere für Daten aus Mehrpersonen-Gesprächen, bei denen der Transkriptionsaufwand besonders hoch ist. Das Institut für Deutsche Sprache in Mannheim baut seit geraumer Zeit Gesprächskorpora auf und macht sie der computergestützten Recherche zugänglich. Aber während die an das Recherchesystem COSMAS-II angeschlossenen schriftlichen Korpora inzwischen knapp zwei Milliarden laufende Wörter enthalten,[1] liegt das recherchierbare Gesprächskorpus bei etwas über zwei Millionen laufenden Wörtern. Das ist allerdings im Vergleich mit anderen Korpora gesprochener Sprache schon beachtlich (vgl. Kallmeyer 1997).

Zur Gesamtarchitektur korpustechnologischer Werkzeuge, auf die sich der Ausbau des Gesprächskorpous im IDS stützt, gehören Transkriptionseditoren (wie DIDA[2] oder in Zukunft auch EXMARaLDA[3]), die Digitalisierung von analogen Aufnahmen, das Alignment als Synchronisation von Transkript und Sprachsignal der Gesprächsaufnahme (vgl. Schmidt/Neumann 1999), eine Verwaltungsdatenbank für Metainformationen zu den Korpora und eine Anwendung für Gesprächsdaten des am IDS entwickelten Volltext-Recherchesystems COSMAS II (vgl. Bodmer/Fach/Schmidt/Schütte 2002). Wir konzentrieren uns hier auf den zuletzt genannten Punkt, beschreiben das Modell für Diskurstranskripte sowie die Suchoperatoren in COSMAS II und führen Ergebnisse einer Recherchesequenz vor.

1. Das Diskurstranskriptmodell in der Volltextdatenbank COSMAS II

Aus den Grundmerkmalen von Gesprächstranskripten lässt sich ein Diskurstranskriptmodell ableiten. Durch die technische Umsetzung des Modells in COSMAS II werden diskursanalytische Recherchen möglich, die herkömmliche Volltextdatenbanken nicht bieten können. Das Diskurstranskriptmodell umfasst folgende Merkmale:

---

[1] 2002 umfasste das Korpusarchiv 1,9 Milliarden Textwörter (vgl. http://www.ids-mannheim.de/kt/projekte/korpora/archiv.html).

[2] http://www.ids-mannheim.de/prag/dida/.

[3] http://www.rrz.uni-hamburg.de/exmaralda/.

Der Text eines Transkripts besteht aus einem Redestrang pro Gesprächsteilnehmer. Jeder dieser Stränge muss über Unterbrechungen, sei es durch kurze Einschübe oder durch längere Pausen des Teilnehmers, hinweg verfolgt werden können.

Simultanpassagen müssen eine Anfangs- und Endbegrenzung haben. In Simultanpassagen können Wörter unterschiedlicher Sprecher zeitlich nicht in Relation zueinander gebracht werden, sondern werden insgesamt als gleichzeitig gesprochen behandelt.

Änderungen in Lautstärke und Schnelligkeit werden paarweise, d.h. am Anfang und Ende markiert. In manchen anderen Transkriptionssystemen wie GAT ist nur die Markierung des Beginns obligatorisch (vgl. Selting et al. 1998), aber im Hinblick auf die Recherchemöglichkeiten sind paarige Markierungen sinnvoller, auch wenn in der Regel die Schlussmarkierung weniger verlässlich ist als die Markierung des Beginns.

Die kleinste Einheit ist das Wortfragment. In der Regel fällt ein Wort mit einem Fragment zusammen, außer wenn ein Wort durch eine Mikropause oder die Grenze einer Simultanpassage in mehrere Fragmente segmentiert wird.

Ereignisse wie Pausen und nicht lexikalisierte Äußerungen sind eigenständige Einheiten des Diskurses. Im Gegensatz zu anderen, z.B. prosodischen Annotationen beziehen sie sich nicht direkt auf Äußerungen oder Äußerungsteile.

Zwischen den Wörtern eines Sprechers lässt sich ein Wortabstand definieren. Benachbarte Wörter haben einen Wortabstand von 1.

Zwischen den Wörtern, Ereignissen und nicht-lexikalisierten Äußerungen wird eine zeitliche Abfolgerelation definiert, so dass man zu je zwei Einheiten zweier Sprecher sagen kann, ob sie nacheinander oder gleichzeitig vorkommen.

In einem sog. Indexierungsprozess werden die Diskurstranskripte in komprimierter Form in die Datenbank aufgenommen und auf einen schnellen Recherchezugriff vorbereitet. Parallel dazu wird ein Archiv der mit den Transkripten alignierten Audiodateien aufgebaut, um zu Recherchetreffern die Aufnahmestücke abspielen zu können.

2. Suchoperatoren der COSMAS-Gesprächsrecherche

COSMAS II stellt Suchoperatoren speziell für das Recherchieren in Diskurstranskripten zur Verfügung:

Der Wort-Operator *WORT(x)* sucht nach einem oder mehreren alternativen Suchbegriffen, die mit dem Platzhalteroperator * (Asterisk) versehen sein können. So findet *WORT(indirekt*)* Vorkommnisse von *indirekt*, *indirekte*, *indirekter*, *indirektes*, etc. Falls die Großschreibung in den Transkripten verwendet wird, kann durch entsprechende Optionen gesteuert werden, ob die Groß-/Kleinschreibung berücksichtigt werden soll. Durch die Trennung von Text und Annotationen in COSMAS II ist dieser Operator im Stande, Wörter zurückzuliefern, welche durch die Diskursgliederung oder Annotationen fragmentiert sind. Die Suchanfrage *WORT(indirekt)* findet demnach auch Textstellen wie *indi|rekt* mit einer Überlappungsgrenze, *indi*rekt* mit Mikropause oder *indirekt*↓ mit Tonfallmarkierung.

Operatoren für die Suche nach Wörtern mit prosodischen Merkmalen: So ist *BETONUNG* ein Suchoperator, der Wörter findet, bei welchen auffallende Betonungen markiert wurden, z.B. *i"ndirekt* mit auffallend betonter Silbe *in*. Die Operatoren *DEHNUNG* und *DEHNUNG-LANG* finden Wörter mit einer auffallenden Dehnung. *INTONATION(x)* sucht nach

Wörtern mit aufsteigender, absteigender oder schwebender Intonation; z.B. liefert *INTONATION(fallend)* Textstellen wie *indirekt*↓.

Mit dem Operator *GLEICH(X,Y)* lassen sich Wortformen und prosodische Merkmale kombinieren, z.B. das Wort *nicht* sowohl mit steigender Intonation als auch mit einer besonderen Dehnung. Als Treffer liefert das Verfahren ein Textstück wie *willst du ni::cht*↑, übergeht jedoch Fälle wie *glaubst du nicht*↑.

Operatoren für Pausen verschiedener Länge: *PAUSE* für Pausen beliebiger Dauer, *PAUSE-KURZ* für Mikropausen (in unseren Diskurstranskripten ca. ½ Sek.) und *PAUSE-LANG* für Pausen längerer Dauer ab 1 Sek. Diese drei Pausen-Operatoren lassen sich grundsätzlich je nach Art und Genauigkeit der in den Transkripten vorkommenden Pausenkodierungen beliebig erweitern.

Operatoren für nicht-lexikalisierte Äußerungen: Nicht-lexikalisierte Äußerungen bilden in unseren Transkripten eine offene Menge von kurzen Beschreibungen in der Sprecherzeile, z.B. *und als ich sah dass er lacht LACHT KURZ sachte ich*. Diese Beschreibungen sind mit Großbuchstaben festgehalten, damit der Leser sie als Annotation erkennen kann. In COSMAS II werden sie in den Annotationsstrukturen gespeichert und über eine Reihe von speziellen Suchoperatoren recherchierbar gemacht, z.B. über den *LACHT*-Operator, der Beschreibungen wie *LACHT*, *LACHT KURZ*, *LACHANSATZ*, etc. wiederfindet. Da diese Beschreibungen nicht ausschließlich geschlossene Listen von Ausdrücken verwenden, sondern auch aus freiem Text bestehen können, haben wir einen Operator *VOCAL(x)* eingeführt, der mit freiem Text als Eingabe arbeitet, z.B. *VOCAL(LACHANSATZ)*, *VOCAL(GRINST)*, *VOCAL(SEUFZT)*.

Operatoren für die Diskursgliederung liefern Wörter zurück, die sich an markanten Stellen der Diskursgliederung wie Äußerungsgrenzen oder Simultanpassagen befinden. So liefern *ÄUßER-BEG* und *ÄUßER-END* jeweils das erste bzw. letzte Wort einer Äußerung zurück und *SIM-BEG* bzw. *SIM-END* jeweils das erste bzw. letzte Wort einer Simultanpassage.

Operatoren für Sprecherdaten wie Altersangabe oder Geschlecht, die mit den anderen Operatoren kombinierbar sind. *IN-SP-GESCHL(X Y)* filtert die Suchanfrage X auf Sprecher mit Geschlecht Y; *IN-SP-ALTER(X Y)* filtert die Suchanfrage X auf Sprecher mit Alter Y. Bislang enthalten nur wenige Gesprächstranskripte diese Informationen, COSMAS II ist aber darauf vorbereitet, diese Daten bei Suchanfragen berücksichtigen zu können.

Ein sprecherbezogener Wortabstandsoperator *ABSTAND(X nw Y)* gewährleistet, dass die gefundenen X- und Y-Textstellen vom selben Sprecher stammen und höchstens den angegebenen Abstand (in Worteinheiten) aufweisen. Dieser Operator kann X- und Y-Stellen miteinander kombinieren, zwischen denen Einschübe von anderen Sprechern vorkommen. So liefert *ABSTAND(WORT(nicht) 3w WORT(weil))* *nicht-weil*-Wortpaare eines Sprechers zurück, die höchstens 3 Wörter auseinander liegen; *ABSTAND(WORT(nicht) 0w ÄUßER-END)* liefert *nicht*-Stellen am Ende einer Äußerung zurück; und. *ABSTAND(WORT(nicht) +1w INTONATION(steigend))* liefert *nicht*-Stellen zurück, die unmittelbar vor einem beliebigen Wort (*+1w*) mit steigender Intonation gesprochen wurden.

Der Abfolgeoperator *ABSTAND(X ns Y)* betrachtet das Transkript als Abfolge von Zeitsegmenten. Eine Simultanpassage gilt in dieser Betrachtungsweise als ein Zeitsegment. Wörter außerhalb von Simultanpassagen bilden jedes für sich ein Segment. Im Gegensatz zum Wortabstand stellt dieser Abstandsoperator sowohl für Wörter eines einzelnen Sprechers als auch verschiedener Sprecher einen Bezug her. Er ist insbesondere als Instrument

zur Untersuchung von Phänomenen in und um Simultanpassagen wichtig. Mit dem Abfolgeoperator können nicht-lexikalisierte Äußerungen berücksichtigt werden, die der Wortabstandsoperator außer Acht lässt. So sucht die komplexe Anfrage *ABSTAND(IN-SP-GESCHL(WORT(weil) fem) +1s IN-SP-GESCHL(WORT(nein nee nö) masc))* nach Vorkommnissen von *weil* in Äußerungen von Sprecherinnen, denen unmittelbar (+1s) danach ein *nein* (oder *nee* oder *nö*) eines männlichen Teilnehmers folgt, also z.B.:

- AA(♀): das stimmt weil | ich |
- BB(♂): |nee nee| das können sie so nich sagen

oder

- AA(♀): das stimmt weil↓
- BB(♂): nein das können sie so nich sagen

3. Zur Arbeitsweise der COSMAS-Gesprächsrecherche

Beim Recherchieren geht man in folgenden Schritten vor:

Transkriptauswahl. Man hat die Wahl zwischen dem gesamten Datenbankinhalt, einem vordefinierten Korpus (in der Regel einer Sammlung von Transkripten, die für ein Projekt zusammengestellt worden sind) oder einem virtuellen Korpus, das man sich selbst zusammenstellt. Als Auswahlkriterien stehen dabei zur Verfügung: das Aufnahmedatum (wenn bekannt), die Textsorte (eine Bezeichnung für die während der Aufnahme vorherrschende Art des Diskurses: *Interview*, *Schlichtung*, *Talk Show*, usw.), die Korpusbezeichnung und Textfelder für frei verfasste Beschreibungen oder eine Liste von Schlüsselwörtern. Diese Kriterien lassen sich untereinander kombinieren.

Recherchieren. Die Suchoperatoren werden einzeln oder in Kombination verwendet. So lassen sich Wortformen, Prosodie, Diskursgliederung und Sprecherdaten zu Suchanfragen kombinieren. Frühere Suchanfragen können in neue Anfragen eingefügt werden oder abgespeichert werden, um sie in späteren Sitzungen wieder zu verwenden. Bei der Wortformsuche kann man Platzhalteroperatoren einsetzen. Diese erzeugen Auswahllisten, aus welchen man die unerwünschten Wortformen abwählen kann.

Ergebnispräsentation: Die primäre Darstellungsform der Treffer ist das KWIC („Key Word in Context" als einzeilige Kurzform des Wortlauts des Treffers ohne weitere Informationen aus dem Transkript). Diese Darstellung lässt sich nach verschiedenen Kriterien (z.B. alphabetisch oder chronologisch) sortieren. Die Trefferstatistik kann nach verschiedenen Gesichtspunkten aufgelistet werden: z.B. zusammengefasst nach Korpuszugehörigkeit, Diskurstyp oder Aufnahmedatum. Vom KWIC aus gelangt man in die Partiturdarstellung, die den Treffer in einem größeren Interaktionskontext präsentiert. Sowohl im KWIC als auch in der Partiturdarstellung kann man die Treffer bei alignierten Diskursen in einem frei wählbaren Zeitfenster abspielen. Das KWIC kann im ASCII-Format zum Zwecke der automatischen Weiterverarbeitung oder im RTF-Format exportiert werden; in der nächsten Programmversion wird auch die Partiturdarstellung zu exportieren sein. Einzelne Treffer aus alignierten Transkripten stehen als Audio-Ausschnitte für die weitere Signalverarbeitung (z.B. in Praat) zur Verfügung.

4. Eine exemplarische Recherchesequenz in COSMAS II zur Distribution von *aber*

Die Recherchemöglichkeiten in COSMAS II lassen sich durch eine Sequenz modifizierter Suchanfragen zur Konjunktion *aber* demonstrieren. Ziel dabei ist zum Einen, typische Verteilungsmuster zu erkennen: Mit welchen Wörtern in welchem Abstand und mit welchen prosodischen Markierungen wird *aber* kombiniert, an welchen Stellen im interaktiven Geschehen taucht es auf? Zum Anderen geht es darum, in einer Recherchesequenz notwendige Zusatzbedingungen, durch die sich Treffer zur Konjunktion *aber* von anderen Fällen (z.B. *aber* als Modalpartikel) unterscheiden lassen, als restriktive Filter zu erkennen und zu prüfen.

Zunächst suchen wir nach *aber* im Rahmen einer einfachen Wortsuche; das ergibt – in einem kleinen Korpus von bereits text-ton-synchronisierten 15 Diskursen – unübersichtliche 1281 Treffer. Diese Anfrage wird angereichert durch die Einbeziehung von *aber* mit Aussprachevarianten; da ein Lemmatisierer noch nicht zur Verfügung steht, müssen die möglichen Notationsvarianten explizit in der Anfrage genannt werden: *abba aba awwa awa.* Eine Anfrage beschränkt auf diese dialektal notierten Varianten ergibt zehn Treffer:

zu eidl ja do haww=isch **awwa** ga ned dro isch hab
isch wirklisch nischt mehr komm **awwa** am samstag sach=isch ehrlisch
bin un donz nix mehr **awwa** do konn=sch misch nimeh setze
ah wie konn do des **awwa** soi des haww isch awwa
awwa soi des haww isch **awwa** vorher nischd gehabt nischt gehabt
ah des sieht ma ihne **awwa** ehrlich net net an ja
ihne in die fress ja **awwa** wonn sie misch schneide fa
(2001.15,gerddlA)
huschde der kummt vum rache **awwa** die leute haben äh gut
(2001.26,zehner)
geh daß isch mer glei=n **awa** wenn se misch noch=emol so
(3001.03,alte-sau)
agestellt woadn was i **awwa** no schlimmer empfunden hab des
(4050.026,abtreibung)

Acht der zehn Treffer stammen aus zwei Interviews aus dem Projekt „Stadtsprache Mannheim", die beiden restlichen aus einer Schlichtungsverhandlung (gleichfalls aus Mannheim) und einer Talkshow, in der eine Beteiligte mit bairischem Dialekt spricht.

Weitere Restriktionen ergeben sich durch Kombination von *aber* mit prosodischen Merkmalen wie Betonung, steigendem Grenzton (also funktional gesehen einer Fortsetzungsintonation), *aber* nach einer Pause oder mit steigendem Grenzton und nachfolgender Pause (das ergibt eine starke Projektion, also Fortsetzungserwartung, die u.a. das Rederecht sichern hilft). Weitere Möglichkeiten für eine Anreicherung der einfachen Anfrage und damit stärkere Filterung der möglichen Treffer sind:

Kombinationen von Verknüpfungsformen wie *zwar – aber* mit unterschiedlichem Wortabstand (bei zunehmendem Wortabstand wächst zwar die Trefferzahl, zugleich steigt aber die Wahrscheinlichkeit, dass die beiden Vorkommen von *zwar* und *aber* syntaktisch nicht zusammengehören) und *ja – aber*; im Unterschied zu *zwar* müssen hier viele Treffer manuell ausgesondert werden, bei denen *ja* z.B. als Modalpartikel oder Rückmeldeelizitierung fungiert.

aber und Interaktivität: bei Sprecherwechsel, zugleich am Beginn von Simultanpassagen (dadurch kann der Folgesprecher Widerspruch ankündigen; diese Fälle können auch daraufhin untersucht werden, ob hier zwei – oder mehr – Sprecher um das Rederecht konkurrieren).

Bei der Ausarbeitung derartiger Recherchesequenzen können wir sehen, dass sich Suchanfragen in COSMAS II oft sinnvoll in einer Art ‚Trial&Error'-Verfahren entwickeln lassen; die erste Formulierung der Suchanfrage wird auf Grund vieler unspezifischer Treffer als ungenau erkannt und kann durch weitere Bedingungen treffsicherer werden. Ein Beispiel dazu ist eine Suchanfrage nach *aber* (mit Aussprachevarianten) am Beginn einer Simultanpassage, mit der nach der Konjunktion am Äußerungsbeginn gesucht werden soll. Schon der erste von 156 Treffern zeigt ein unerwartetes Muster:

Abb. 1: Treffer zu *aber* (Ausschnitt aus Volltextdarstellung)[4]

Hier wird in einem Beratungsgespräch *aber* vom Ratsuchenden (RS) simultan mit dem Rezeptionssignal *mh* des Beraters (BR) geäußert, aber nicht am Beginn der Äußerung. Eine Erweiterung der Suchanfrage durch die Zusatzbedingung ‚Sprecher Beginn' eliminiert diesen Treffer. Ein typischer von den nurmehr 34 Treffern sieht im Partiturformat so aus:

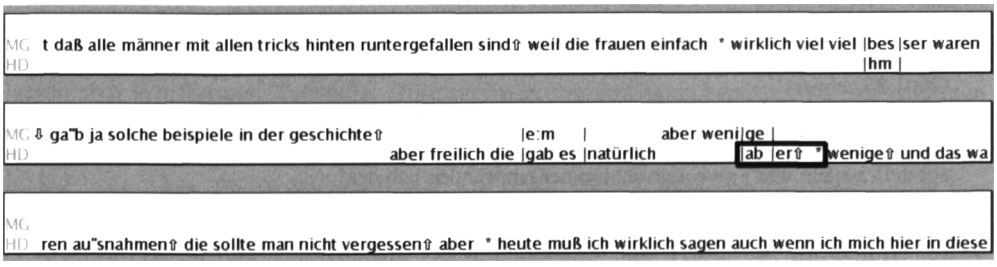


Abb. 2: Treffer zu *aber* (Ausschnitt aus Volltextdarstellung)

Hier kommentiert HD eine Sachverhaltsdarstellung von MG zur gesellschaftlichen Rolle von Frauen, nachdem MG HD mit fragender Intonation zur Rederechtsübernahme eingeladen hat. Nach dem kurzen relativierenden Einwurf von MG (*aber wenige*) setzt HD mit einer kleinen Überlappung (so dass *aber* tatsächlich partiell simultan gesprochen wird) ihren Kommentar fort. Gesprächsanalytisch könnte man ihre beiden Teiläußerungen vor oder nach dem Einwurf als einen Turn interpretieren – COSMAS II geht wegen der Lücke im Sprechtext während MGs Einwurf aber ganz formal von zwei Turns aus.

---

[4] In der COSMAS-II-Partiturdarstellung wird der Treffer blau markiert; zur Verdeutlichung sind hier in den Bildschirmfotos die Treffer durch schwarze Umrahmung gekennzeichnet.

Eine Rechercheanfrage nach *aber* (einschließlich der dialektalen Aussprachevarianten) mit den Zusatzbedingungen mit steigender Intonation (also als Fortsetzungssignal) und vor einer Pause unbestimmter Länge ergibt im Korpus der alignierten Transkripte nur 4 Treffer:

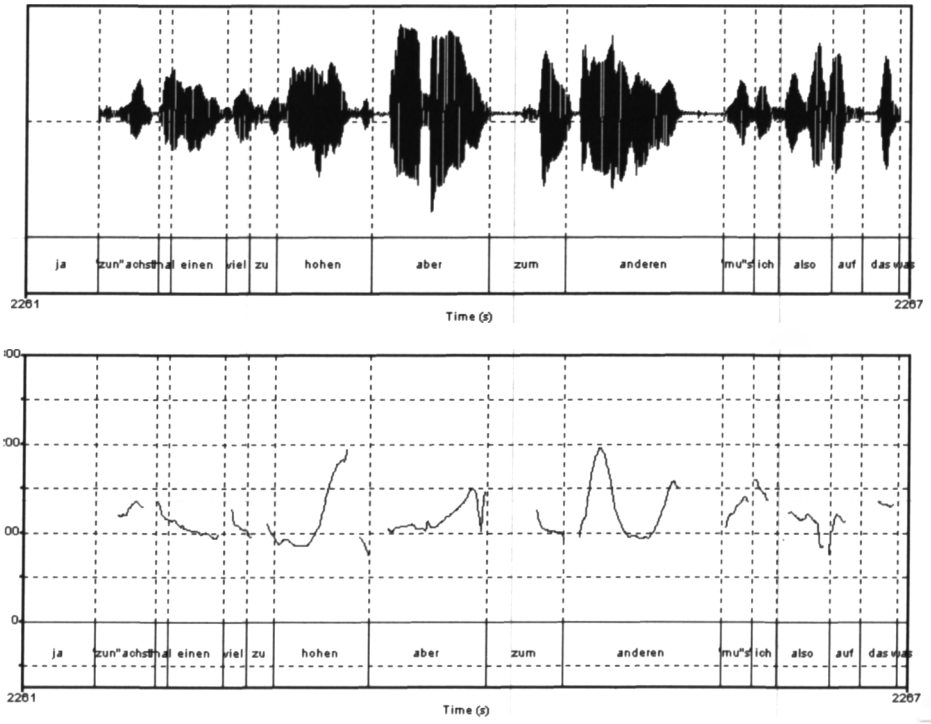
damit schbeziell net zuviel belasten **aber** egal wie=s ausgeht schon die (2001.26,zehner)
mal einen viel zu hohen **aber** zum anderen muß ich also (4050.058,demontage)
dann kommt doch immer ein **aber** rachel was denkst du dazu
(4050.192,wirsindwiederwer)
haben finde ich völlig richtig **aber** unsere oma nicht und dann
(4051.08,muell-gerb)

Mit diesen vier Treffern möchten wir uns nacheinander etwas näher befassen. In der Partiturdarstellung zeigen die Treffer (im Gegensatz zur KWIC-Darstellung) die erfragte Kombination von Wort und Annotationen an – hier der erste Treffer:

Abb. 3: Treffer zu *aber* (Ausschnitt aus Volltextdarstellung)

Interessant ist hier die Folge zweier durch *awwer* bzw. *aber* eingeleiteter Konstruktionen; durch die prosodische Abtrennung des zweiten *aber* entsteht die Frage, ob der Konnektor auf das erste oder das zweite der folgenden Satzkonstruktionen bezogen ist. Hier auch der zweite Treffer ausschnittweise in Partiturdarstellung:

Abb. 4: Treffer zu *aber* (Ausschnitt aus Volltextdarstellung)



Beim Anhören dieses Treffers wird eine auffällige Projektionstechnik deutlich: Der Sprecher intoniert sowohl *hohen* als auch *aber* und *zum anderen* mit einem ausgeprägten Fortsetzungssignal und lässt danach jeweils eine Pause folgen – er sichert sich damit das Rederecht, stuft die nachfolgende Sachverhaltsdarstellung vorgreifend in der Relevanz hoch und verschafft sich zugleich Planungszeit zur Ausformulierung seiner Äußerung. Eine Prosodieanalyse zu diesem Recherchetreffer mit ‚Praat'[5] ergibt folgende grafische Darstellung:

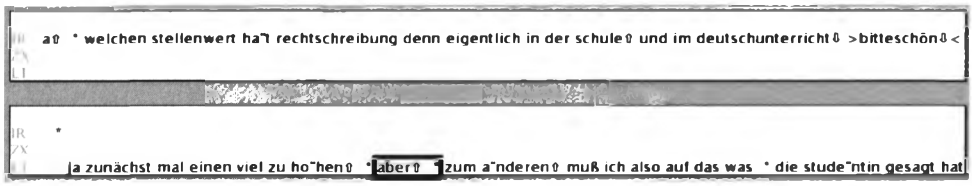


Abb. 5: Prosodieanalyse mit ‚Praat'

Deutlich erkennbar sind in der Waveform die drei Pausen nach *hohen*, *aber* und *anderen*, in der Pitch-Darstellung (also der Grundfrequenz/F0-Kurve) die Tonsprünge, mit über 100 Hz nach *hohen* besonders ausgeprägt, aber auch nach *anderen* und nach *aber* mit über 50 Hz markant; auffällig ist zudem die expressive zweigipflige Kontur bei *anderen* mit einem Tonsprung auf der Akzentsilbe, der vergleichbar weit ausfällt wie die Grenzintonation bei *hohen*. In der Praat-Grafik werden Waveform und Grundfrequenzkurve jeweils mit einem Ausschnitt aus der in Praat importierten Alignment-Labeldatei kombiniert; dadurch werden Segmente der Analysekurven durch Zuordnung zur transkribierten Wortfolge interpretierbar.

Der dritte Treffer (*dann kommt doch immer ein aber rachel was denkst du dazu*) betrifft eine substantivierende Verwendung von *aber*; das zeigt, dass in Abwesenheit morphosyntaktischer Annotationen im Transkript auch *aber*-Treffer möglich sind, die durchaus die prosodischen Zusatzbedingungen erfüllen, aber dennoch keine Konjunktionen betreffen.

Beim vierten Treffer schließlich ist eine interessante Verwendungsvariante von *aber* zu beobachten – in einem ethnografischen Interview mit einem Fernsehmoderator (MG) benutzt dieser für seine Schlussfolgerung *da" müssen sie sehr viel fingerspitzengefühl haben* eine Belegerzählung mit einem fiktiven Zitat (*aber↑ * unsere o"ma↓*), um mit dieser elliptischen Kurzformel einen aus seiner Sicht störenden Typ thematischer Initiativen von Talkgästen anzudeuten; hier finden wir also durchaus das Muster ‚Berücksichtigung + eigene Position', aber nicht in der Gesprächsbeziehung zwischen den hier Beteiligten WS und MG, sondern in einer von MG referierten anderen Gesprächssituation:

[5] Vgl. http://www.praat.org.

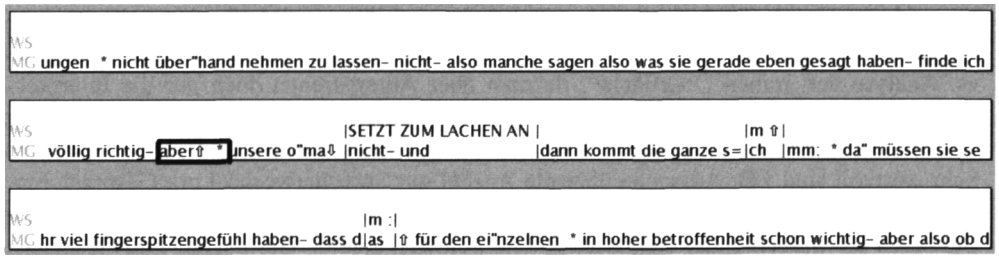


Abb. 6: Treffer zu *aber* (Ausschnitt aus Volltextdarstellung)

5. Ausblick: Möglichkeiten der Analyse von Gesprächskorpora

Eine Recherche wie die oben demonstrierte ermöglicht einen Zugriff auf eine Vielzahl von gleichartigen Belegen aus dem Gesamt-Korpus oder aus beliebig zusammengestellten Teilkorpora. Zudem kann die Recherche auch Angaben zur Verteilung von Ausdrücken, Struktureigenschaften und Quantitäten im Korpus liefern. Ziele der gegenwärtig möglichen Datenbankrecherche in Gesprächskorpora können sein:

> Unterstützung der heuristischen Mustererkennung. Dabei sind die aus der vorgängigen Analyse bekannten bzw. vermuteten Mustereigenschaften in die begrenzten Möglichkeiten der COSMAS-Abfrage zu übersetzen. Die Auseinandersetzung mit den Treffern und ihre Bewertung als Kandidaten für das Muster schärft den Blick für die zu explizierenden Mustereigenschaften und für die Varianz der Musterrealisierungen.
> Verteilungsanalyse, z.B. die Verteilung von Wörtern und Ausdrucksmustern auf Textsorten / Genres.

Grundsätzlich kann man nur automatisch suchen, was im Korpus notiert worden ist. Das DIDA-Gesprächskorpus des IDS enthält Transkriptionen einer mittleren Feinheit. Wie schon oben angedeutet, gibt es Untersuchungsinteressen, für die eine Verfeinerung der Transkriptionsannotation notwendig ist, z.B. genauere prosodische Markierungen, d.h. präzisere Pausen-, Tempo-, Lautstärke- und Intonationsangaben, wie sie das Transkriptionssystem GAT vorsieht. Im Prinzip sind im DIDA-Korpus solche Verfeinerungen in begrenztem Umfang in der Textzeile und ansonsten in Annotationszeilen möglich. So wird für bestimmte Forschungszwecke z.B. die Anreicherung der Wortformen mit ihrer Aussprache (in Lautschrift) wünschenswert (vgl. das IDS-Projekt „Variation des gesprochenen Deutsch" - INTERNETADRESE). Weiter kann man eine Lemmatisierung für die gesprochensprachlichen Wortformen entwickeln – beileibe keine triviale Aufgabe. Die Lemmatisierung kann wiederum die Grundlage für eine morpho-syntaktische Annotation sein. Natürlich wäre es großartig, wenn man in Gesprächskorpora z.B. den Zusammenhang zwischen Wortstellung und Turnstruktur recherchieren könnte; im Verhältnis dazu erscheinen die gegenwärtig realisierten Möglichkeiten der Wortsuche natürlich noch recht einfach. Träumen kann man auch davon, dass über den Anschluss eines Lexikons bestimmte semantische Eigenschaften recherchierbar werden. Bis derart weit entwickelte Werkzeuge zur Verfügung stehen, ist es allerdings wohl noch ein weiter Weg. Für die Entwick-

lungsarbeit ist auch die Auseinandersetzung mit zwei divergenten Strategien der Korpuslinguistik erforderlich: Annotieren, um spezifischere Recherchen durchführen zu können vs. Analyse der ‚reinen Oberfläche', da man über Annotationen doch nur die Interpretationen wieder herausbekommt, die man hineingesteckt hat.

Die Recherchemöglichkeiten sind auch durch den Ausbau von COSMAS zu erweitern. Sehr wertvoll sind z.B. Aussagen über die Anzahl der Sprecherwechsel pro Gespräch bzw. pro Gesprächseinheit (zeitlich bestimmte Abschnitte im Gesprächsverlauf), der Anzahl von Beiträgen pro Sprecher, die durchschnittliche Beitragslänge in Gesprächen und Gesprächsphasen, die Länge der Redebeiträge pro Sprecher, über das Lexikon der einzelnen Sprecher und das Lexikon der Interaktion (u.a. als Zugang zur Themen- und Genrespezifik). In einer Folgeversion von COSMAS II sollen statistische Werkzeuge, die am IDS für die geschriebene Sprache angewendet werden, auch auf die Transkripte angesetzt werden, insbesondere eine Kollokationsanalyse, die Abweichungen von Wortfügungen im gewählten Korpus von der statistisch erwartbaren Verteilung in der Grundgesamtheit erfasst und auch von der Verteilung in schriftlichen Korpora. Dann wird es z.B. auch möglich, Indikatoren für Genres und andere Eigenschaften von Interaktionen auf statistischer Grundlage zu bestimmen.

Literatur

Bodmer, Franck / Fach, Marcus L. / Schmidt, Rudolf / Schütte, Wilfried (2002): Von der Tonbandaufnahme zur integrierten Text-Ton-Datenbank. Instrumente für die Arbeit mit Gesprächskorpora. In: Pusch, Claus D. / Raible, Wolfgang (Hg.): Romanistische Korpuslingustik: Korpora und gesprochene Sprache. Romance Corpus Linguistics: Corpora and Spoken Language. Tübingen. (ScriptOralia 126), S. 209-243.

Kallmeyer, Werner (1997): Vom Nutzen des technologischen Wandels in der Sprachwissenschaft: Gesprächsanalyse und automatische Sprachverarbeitung. In: Zeitschrift für Literaturwissenschaft und Linguistik 107, S. 124-149.

Schmidt, Rudolf / Neumann, Robert (1999): Automatic Text-to-Speech-Alignment: Aspects of Robustification. In: Matousek, V. / Mautner, P. / Ocelíková, J. / Sojka, P. (Hg.): Text, Speech and Dialogue (Lecture Notes in Artificial Intelligence). Berlin/Heidelberg. S. 72-76.

Selting, Margret / Auer, Peter / Barden, Birgit / Bergmann, Jörg / Couper-Kuhlen, Elisabeth / Günthner, Susanne / Quasthoff, Uta / Meier, Christoph / Schlobinski, Peter / Uhmann, Susanne (1998): Gesprächsanalytisches Transkriptionssystem (GAT). In: Linguistische Berichte 173, S. 91-122.